Homework 1

Handed out: Tuesday, January 22, 2019

Due: Thursday, January 31, 2019 Midnight

Notes:

- We *highly* encourage typed (Latex or Word) homework. Compile as single report containing solutions, derivations, figures, etc.
- Submit all files including report pdf, report source files (e.g. .tex or .docx files), data, figures produced by computer codes and programs files (e.g. .py or .m files) in a **.zip** folder. Programs should include a Readme file with instructions on how to run your computer programs.
- Zipped folder should be submitted to ai.at.cics.nd@gmail.com. One should follow the following naming scheme:
  HW1_LastName_FirstName.zip
- Collaboration is encouraged however all submitted reports, programs, figures, etc. should be an individual student's writeup. Direct copying could be considered cheating.
- Homework problems that simply provide computer outputs with no technical discussion, Algorithms, etc. will receive no credit.
- Software resources for this Homework set can be downloaded from this link.

1 Beta updating from censored likelihood and MLE for uniform distribution

A. Suppose we toss a coin $n = 5$ times. Let X be the number of heads. Assume that we observe that there are fewer than 3 heads, however, we don't know the exact number. We consider the prior probability of heads to be $p(\theta) = \text{Beta}(\theta|1, 1)$. Compute the posterior $p(\theta|X < 3)$ up to normalization constant (i.e., derive expression proportional to $p(\theta|X < 3)$).

B. Consider a uniform distribution centred on 0 with width $2a$. The density function is given as

$$p(x) = \frac{1}{2a}\mathbb{I}(x \in [-a, a]) \tag{1}$$

(a) Given a dataset $x_1, \ldots, x_n$, what is the MLE of a?

(b) What probability would the model assign to $\hat{x}_{n+1}$ using the MLE estimate of a?

(c) Do you see any problem with the above approach? If yes, briefly suggest a better alternative.

2 Bayesian analysis for Uniform distribution and the Taxicab problem

A. Consider the uniform distribution $\mathcal{U}(0,\theta)$. Given the Pareto prior, the joint distribution of θ and $\mathcal{D} = (x_1, \ldots, x_N)$ is

$$p(\mathcal{D}, \theta) = \frac{Kb^K}{\theta^{N+K+1}} \mathbb{I}(\theta \geq \max(\mathcal{D})) \tag{2}$$

Let $m = \max(\mathcal{D})$. The evidence (the probability that all N samples came from the same uniform distribution) is

$$\begin{aligned} p(\mathcal{D}) &= \int_m^\infty \frac{Kb^K}{\theta^{N+K+1}} d\theta \\ &= \begin{cases} \frac{K}{(N+K)b^N} & \text{if } m \leq b \\ \frac{Kb^K}{(N+K)m^{N+K}} & \text{if } m > b \end{cases} \end{aligned} \tag{3}$$

Derive the posterior $p(\theta|\mathcal{D})$ and show that it can be expressed as a Pareto distribution.

B. Suppose you arrive in a new city and see a taxi numbered 10. How many taxis are there in this city? Let us assume taxis are numbered sequentially as integers starting from 0, up to some unknown upper bound θ. Hence the likelihood function is $p(x) = \mathcal{U}(0,\theta)$.

(a) Suppose we see one taxi numbered 100. Using an non-informative prior on θ of the form $p(\theta) = Pa(\theta|0,0) \propto \frac{1}{\theta}$, what is the posterior $p(\theta|\mathcal{D})$?

(b) Compute the posterior mean, mode and median number of taxis in the city, if such quantities exist.

(c) Rather than trying to compute a point estimate of the number of taxis, we can compute the predictive density over the next taxicab number using

$$p\left(D'|\mathcal{D}, \alpha\right) = \int p\left(D'|\theta\right) p(\theta|\mathcal{D}, \alpha)\, d\theta = p\left(D'|\beta\right), \tag{4}$$

where $\alpha = (b, K)$ are the hyperparameters and $\beta = (c, N+K)$ are the updated hyperparameters. Now consider the case $\mathcal{D} = \{m\}$, and $D' = \{x\}$. Using Part A of this problem, write down an expression for $p(x|\mathcal{D}, \alpha)$. As above, consider, $b = K = 0$.

(d) Use the predictive density formula to compute the probability that the next taxi you will see (say, the next day) has number 100, 50 or 150.

(e) Briefly describe (1-2 sentences) some ways we might make the model more accurate at prediction

3 Naive Bayes

A. Consider a Naive Bayes model (multivariate Bernoulli version) for spam classification with the vocabulary V="secret", "offer", "low", "price", "valued", "customer",

"today", "dollar", "million", "sports", "is", "for", "play", "healthy", "pizza". We have the following example spam messages "million dollar offer", "secret offer today", "secret is secret" and normal messages, "low price for valued customer", "play secret sports today", "sports is healthy", "low price pizza". Give the MLEs for the following parameters: θ_{spam}, $\theta_{\text{secret}|\text{spam}}$, $\theta_{\text{secret}|\text{non-spam}}$, $\theta_{\text{sports}|\text{non-spam}}$, $\theta_{\text{dollar}|\text{spam}}$

B. Write a code for a Naive Bayes classifier (both training and prediction) and use it for the two data sets provided at this link. Both the data sets have four variables namely, `Xtrain`, `Xtest`, `ytrain` and `ytest`. Train the Naive Bayes classifier based on `Xtrain` and `ytrain` and test it based on `Xtest` and `Ytest`. Report the misclassification rate for both the data sets.

4 Emperical Bayes

Consider the problem of predicting cancer rates in various cities. In particular, suppose we measure the number of people in various cities, N_i, and the number of people who died of cancer in those cities, x_i. We assume $x_i \sim \text{Bin}(N_i, \theta_i)$, and we want to estimate the cancer rates θ_i. One way is to estimate the parameters separately, but this will inevitably suffer from the sparse data problem. On the other hand, parameter tying for this problem will be too strong an assumption. A compromise is to assume that the parameters θ_i are similar but there may be city-specific variations. This can be modelled by assuming that the θ_i are drawn from some common distribution. Assuming that $\theta_i \sim \text{Beta}(a, b)$, the joint distribution can be written as:

$$p(\mathcal{D}, \theta, \eta|N) = p(\eta) \prod_{i=1}^{N} \text{Bin}(x_i|\theta_i, N_i) \, \text{Beta}(\theta_i|\eta), \tag{5}$$

where $\eta = (a, b)$. With this setup, the problem can be solved using two methods. In the first method, we can adopt a fully Bayesian approach by assigning an appropriate prior. However, with this setup, we will have to resort to some sampling based approach which will be covered later in the course. The other method is to estimate point estimates for η by using $\mathcal{D}$. This method is known as the empirical Bayes. Note that this method violates the basic assumption of Bayesian statistics as the prior depends on the data.

Write a computer code for estimating the parameters, θ_i by using empirical Bayes. For estimating the parameters η is empirical Bayes, you can utilize the fixed point method proposed by Minka. The MATLAB code for the fixed point method and the cancer data set can be found at this link. The data set has a structure `data` which, in turn, has two variables `y` (No. of persons with cancer) and `n` (Total population of the city).

5 Reject option in classifiers and Newsvendor problem

A. In many classification problems one has the option either of assigning x to class j or, if you are too uncertain, of choosing the reject option. If the cost for rejects is less

than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i =, \ldots, C+1$ where C is the number of classes and $C+1$ is the rejection option. Let $Y = j$ be the true (but unknown) state of nature. We define the loss function as

$$\lambda\left(\alpha_i | Y = j\right) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C+1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{6}$$

In other words you incur 0 loss if you correctly classify, you incur λ_r loss if you choose the reject option and you incur λ_s loss if you make a substitution error (misclassification).

(a) Show that the minimum risk is obtained at $Y = j$ if

$$p\left(Y = j | x\right) \geq p\left(Y = k | x\right) \ \forall k$$

and

$$p\left(Y = j | x\right) \geq 1 - \frac{\lambda_r}{\lambda_s};$$

otherwise we decide to reject.

(b) Describe qualitatively what happens as λ_r / λ_s is increased from 0 to 1 (i.e., relative cost of rejection increase).

B. Suppose you are trying to decide how much quantity Q of some product (e.g., newspapers) to buy to maximize your profits. The optimal amount will depend on how much demand D you think there is for your product, as well as its cost to you C and its selling price P. Suppose D is unknown but has pdf $f\left(D\right)$ and cdf $F\left(D\right)$. We can evaluate the expected profit by considering two cases: if $D > Q$, can evaluate the expected profit by considering two cases: if $D > Q$, then we sell all Q items, and make profit $\pi = \left(PC\right) Q$; but if $D <$Q, we only sell D items, at profit $\left(PC\right) D$, but have wasted $C\left(Q - D\right)$ on the unsold items. So the expected profit if we buy quantity Q is

$$\mathbb{E}_\pi = \int_Q^\infty \left(P - C\right) Q f\left(D\right) dD + \int_Q^\infty \left(P - C\right) D f\left(D\right) dD + \int_Q^\infty C\left(Q - D\right) f\left(D\right) dD \tag{7}$$

Show that the optimal quantity Q^* (which maximizes the expected profit) satisfies

$$F\left(Q^*\right) = \frac{P - C}{P} \tag{8}$$